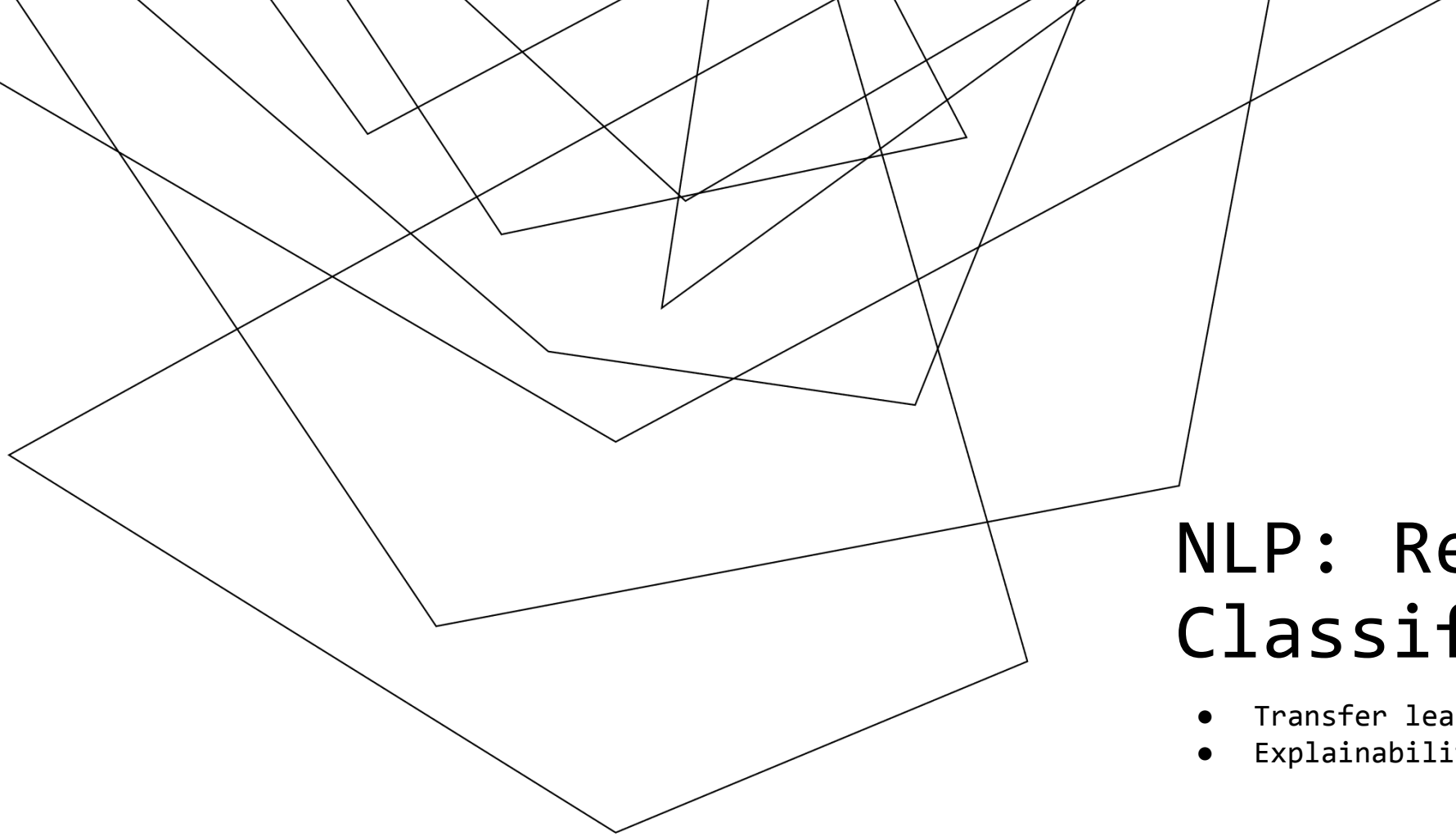


NLP: Relation Classification

- Transfer learning
- Explainability

Alexander Sifel (01427034),
Cowboy (01427034),
Ghost ()

Relation extraction dataset

- Available
  - FoodDisease
  - CrowdTruth Medical Relation Extraction
- Difficulty
  - CrowdTruth dataset has very bad quality
  - Almost all **is_treat** instances are wrongly labeled
  - Linear models unable to converge
- Approach
  - Use only FoodDisease
  - Replace entities with placeholders (term 1: **influence**, term2: **condition**)
  - Calculate Shortest Dependency Path (SDP) for comparison
  - In total 588 rows after preprocessing (132 **is_cause**, 313 **is_treat**)
  - Keep same 10% of samples for testing to make results comparable

Relation (multi-label) classification

- Baseline
  - BoW Naive Bayes classifier

BoW + NB	Precision	Recall	F1	Support
is_cause	1.00	0.07	0.13	14
is_treat	0.81	0.88	0.84	33
micro_avg	0.81	0.64	0.71	47
macro_avg	0.90	0.48	0.49	47

- Improvement ideas
  - BERT features + traditional model
  - Finetuned classifier with BERT encoder

- Choosing BERT
  - Training data should have similar domain as ours: Medicine, Biology
  - We use the **emilyalsentzer/Bio_ClinicalBERT** checkpoint
  - https://arxiv.org/abs/1904.03323

BERT Features + Linear SVC

Features from sentences

- Create padded/truncated token IDs from sentence with tokenizer and encode with **BERT**
- Get last hidden state of the **CLS** special token embedding (first embedding of output sequence)
- For SDP and full sentence

Train model

- Linear SVC
- C: [0.01, 0.1, 1]
- 10-fold cross validation

BERT Features + Linear SVC - results

- Full sentence input better than SDP (same as with BoW + NB)
- Balances precision/recall for **is_cause** (BoW + NB has no recall)
- Accuracy similar to BoW + NB
- Bit higher F1 but much less precision

BERT + SVM	Precision	Recall	F1
is_cause	0.57	0.86	0.69
is_treat	0.86	0.76	0.81
micro_avg	0.74	0.79	0.76
macro_avg	0.72	0.81	0.75

BERT + SVM (SDP)	Precision	Recall	F1
is_cause	0.47	0.50	0.48
is_treat	0.76	0.67	0.71
micro_avg	0.66	0.62	0.64
macro_avg	0.61	0.58	0.60

Finetuned BERT classifier

- Use full sentences (want to learn full context)
- BERT base **CLS** embedding into linear layer with 768 inputs and 2 outputs
- Unlock BERT layer gradients for finetuning
- Multi-label classification -> binary cross-entropy loss
- Early stopping with patience monitoring validation loss
- 10% of train set for cross validation

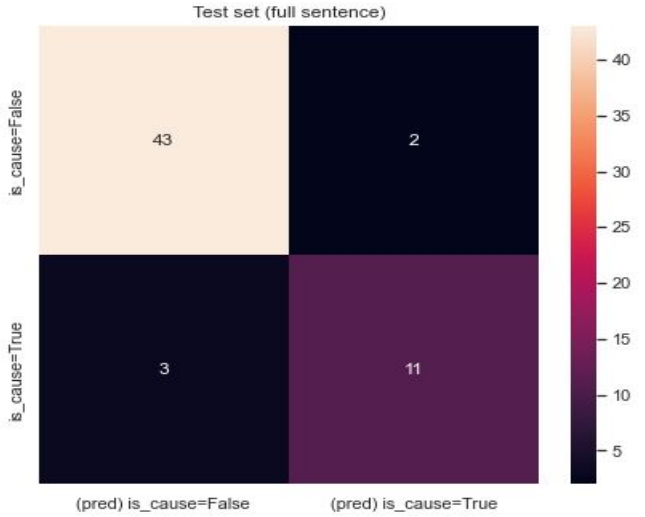
Finetuned BERT classifier - results

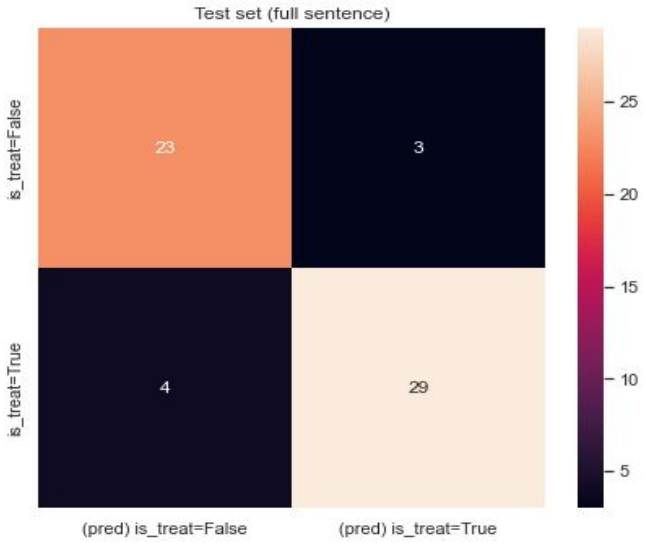
- Finetuning delivers best results by far
- Some (tricky) test examples predictions are still predicted wrongly
- Precision is slightly better than baseline, with much improved recall
- Very high F1 -> try to trade for extra precision by increasing classification threshold

Finetuned	Precision	Recall	F1
is_cause	0.85	0.79	0.81
is_treat	0.91	0.88	0.89
micro_avg	0.89	0.85	0.87
macro_avg	0.88	0.83	0.85

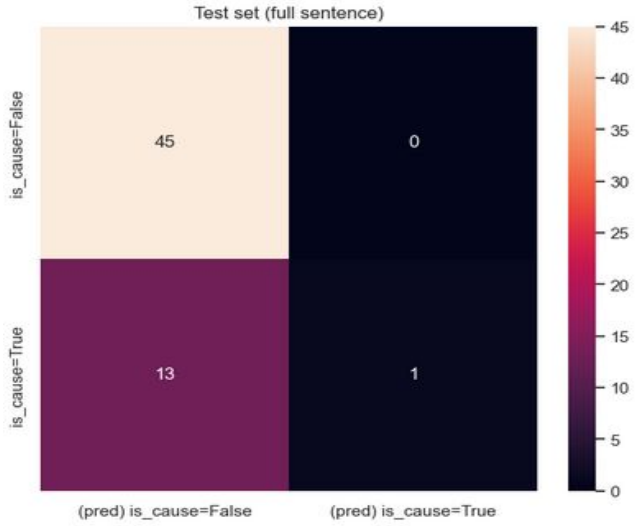
Finetuned BERT Classifier - results

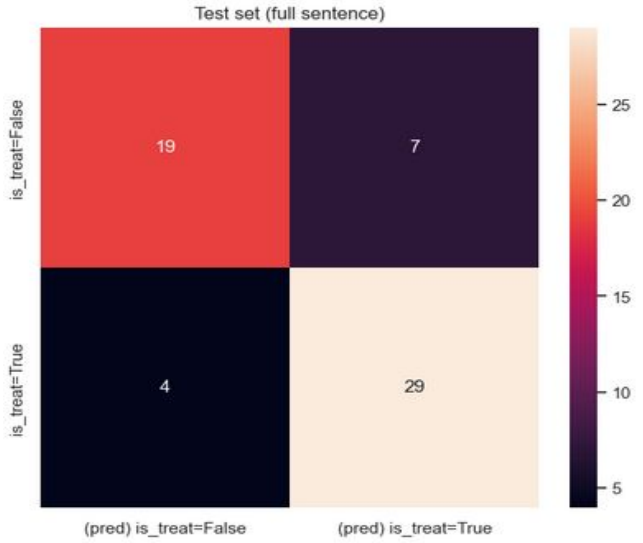
Finetuned





BoW + NB





Finetuned BERT Classifier - results

is_cause false positives:

1. *since influence has been related to the development of chronic condition prevalent in the western world, the use of sweeteners has gradually increased worldwide over the last few years.*

   -> mislabeled, predicted rightly

2. *condition (brd) is a major cause of morbidity and mortality in influence cattle.*

   -> has all the parts, but the relation direction is not right

is_treat false positives:

1. *however, the validity of influence as a treatment for condition (ra), an autoimmune disorder, has not been confirmed yet*

   -> confusion by counterfactual phrasing

2. *abundant studies have highlighted the protective effects of docosahexaenoic acid (dha), in the form of glycerolipids (glycerophosphatides and triglycerides) and dha-ethyl esters (dha-ee) in condition (ad); however, influence (epa) has rarely been implicated*

   -> confusion by counterfactual phrasing

3. *while influence have been shown to exhibit serious side effects, and bioactive compounds from plant-based functional foods have been demonstrated to be active in the treatment of condition with only minimal side effects.*

   -> relation is in there, but not really related to influence

Summary: Classification

- BERT features make an interesting replacement for BoW
- Finetuning gives generally the best performance
- High precision is important, do not want to pollute knowledge base with false positives -> baseline still very competitive

Precision	BoW + NB	BERT + SVM (SDP)	BERT + SVM	Finetuned
is_cause	1.00	0.47	0.57	0.85
is_treat	0.81	0.76	0.86	0.91
micro_avg	0.81	0.66	0.74	0.89
macro_avg	0.90	0.61	0.72	0.88

F1	BoW + NB	BERT + SVM (SDP)	BERT + SVM	Finetuned
is_cause	0.12	0.48	0.69	0.81
is_treat	0.84	0.71	0.81	0.89
micro_avg	0.71	0.64	0.76	0.87
macro_avg	0.49	0.60	0.75	0.85

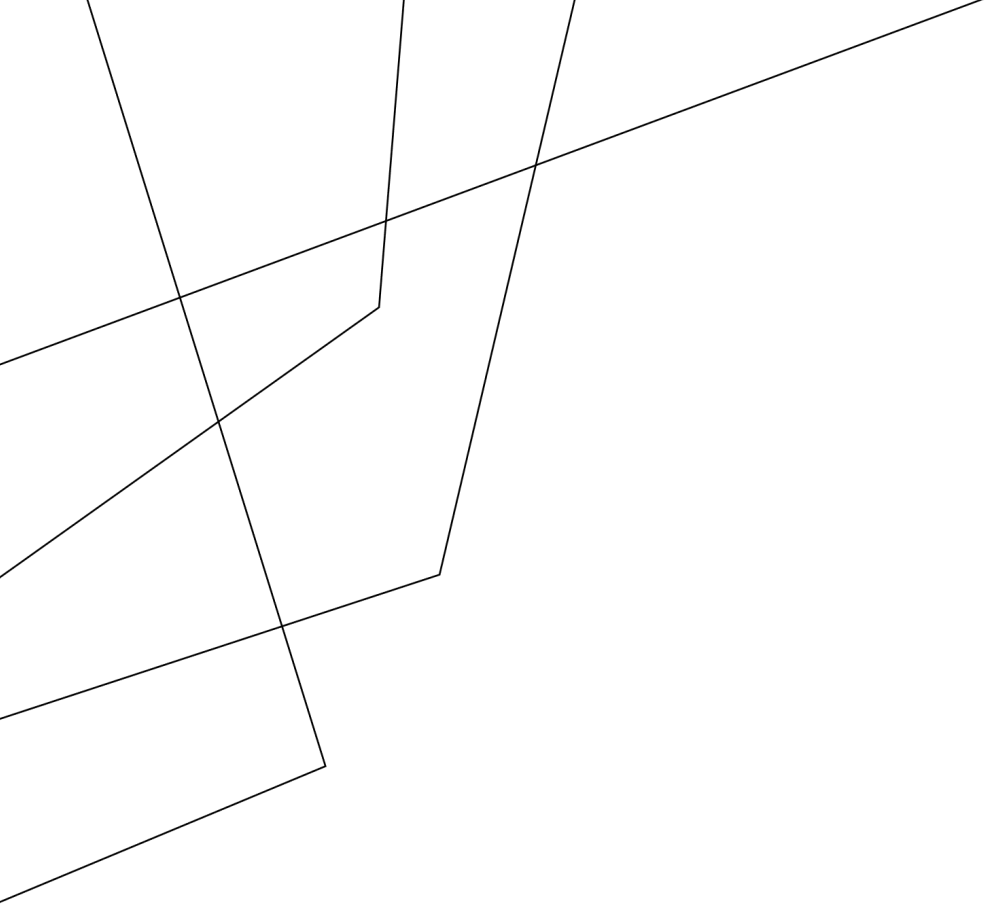
Main Findings & Conclusions

Transfer knowledge

Transfer learning from a BERT model works even with a small dataset

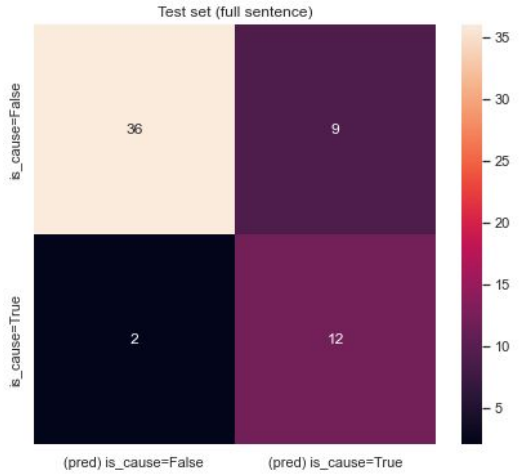
Garbage in - garbage out

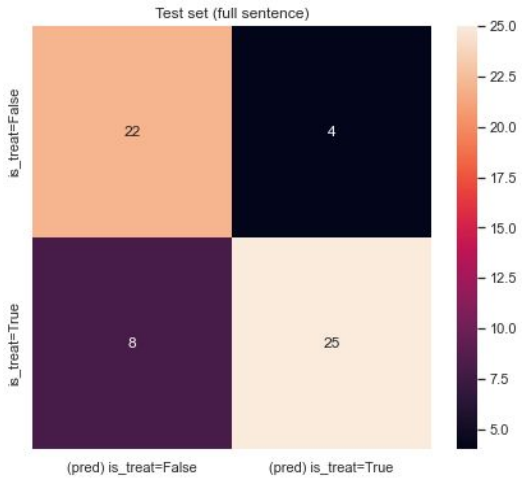
If the dataset is bad enough your model might not converge at all



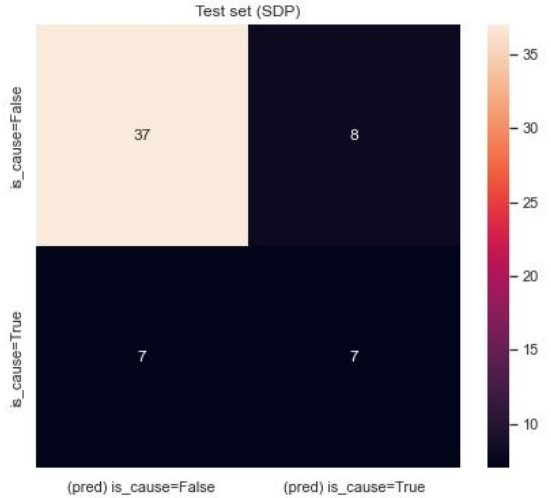
BERT Features + Linear SVC

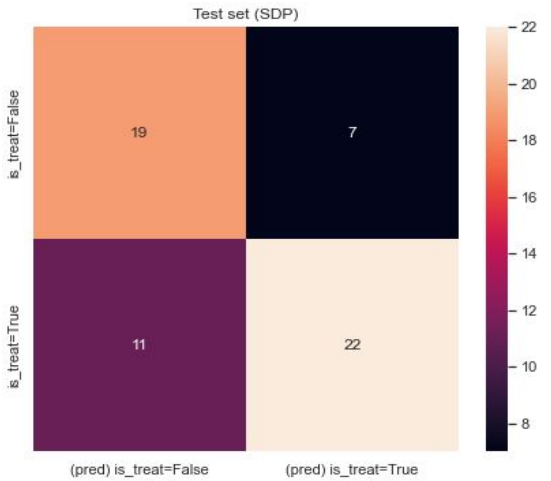
BERT (full)





BERT (SDP)





Baseline (full)